Internet Application

<HW #3>

산업공학과 황영석

2010-12086

1.1 : txt 파일 제출

1.2 : txt 파일 제출

1.3 : 아래 결과 정리

1.3.1 (4class 사용)

This is the total counts  
 ORGANIZATION : 85 PERSON : 144 LOCATION : 77 MISC : 84

1.3.2 (7class 사용)

This is the total counts  
 ORGANIZATION : 67 PERSON : 108 LOCATION : 66 DATE : 39 MONEY : 1 PERCENT : 0 TIME : 0

4class를 사용하였을 떄는, ORGANIZATION, PERSON, LOCATION, MISC 4가지로 분류하였다. 7class를 사용하였을 떄는, ORGANIZATION, PERSON, LOCATION, DATE, MONEY, PERCENT, TIME 7가지로 분류하였다.

두 분류방법을 동일한 Corpus에 대해서 사용하였을 때 결과가 살짝 다르게 나왔다. 예를 들어, 4class의 경우 단체의 갯수를 85개로 측정하였지만 7class는 67개로 측정하였다. 두 분류기가 서로 약간씩은 다르게 학습된 알고리즘을 갖고 NER을 실행하였기에 이런 결과가 나왔을 것이다. 3class를 사용했을 때도 다른 결과가 나왔었는데, 이를 통해서 보건데 NER은 분류기에 따라 성능이 꽤 다를 수 있을 거라 생각했다.

2.1 : txt 파일 제출

2.2 : txt 파일 제출

2.3 : 아래 결과 정리

2.3.1 (english-left3words-distsim.tagger 사용)

{JJ=7157, RB=9159, DT=9253, TO=4138, RP=320, RBR=310, RBS=171, JJS=274, FW=8, JJR=366, NN=14989, NNPS=40, VB=6275, VBN=3718, PDT=306, VBP=2357, PRP=16496, MD=2890, WDT=780, VBZ=1510, WP=862, IN=15213, POS=646, EX=288, VBG=2132, VBD=7873, UH=209, NNS=2904, CC=5036, CD=941, NNP=4904, WRB=775}

2.3.2 (wsj-0-18-left3words-nodistsim.tagger 사용)

{JJ=7131, RB=9228, DT=9317, TO=4138, RP=315, RBR=261, RBS=167, JJS=292, FW=6, JJR=403, NN=15094, NNPS=26, VB=6255, VBN=3743, PDT=297, VBP=2388, PRP=16473, SYM=1, MD=2895, WDT=705, VBZ=1501, WP=899, IN=15210, POS=647, EX=276, VBG=2131, VBD=7879, UH=211, NNS=2903, CC=4984, CD=766, NNP=4983, WRB=775}

english-left3words-distsim.tagger와 wsj-0-18-left3words-nodistsim.tagger를 사용하였을 때 눈에 띄는 차이는 발견할 수 없었다. 각 POS의 숫자는 크게 차이나지 않았으며 기껏해야 전체 corpus에서 몇 개, 혹은 몇 십개의 차이였다. POS tagging 역시 분류방법에 따라 살짝 다른 결과를 보여주며, 이 것이 성능의 차이가 될 수도 있다. (물론 성능은 비슷하지만 잘못 태깅하는 항목이 조금씩 다른 것일수도 있다.)

사소한 다른 점이 있다면, wsj-0-18-left3words-nodistsim.tagger의 경우 SYM이라는 POS가 있었다. 이는 english-left3words-distsim.tagger에는 없는 POS이다. 이는 Symbol에 해당하는 part of speech이다. Pap.txt 파일을 확인해 본 결과 이는 "There \_is\_ something a little stately in him, to be sure," 문장의 is 앞에 있는 \_(underscore)에 해당하였다. 이 문장에서 이게 사용될 이유도 명확하지 않을 뿐더러, underscore가 symbol은 아니므로 이는 오태깅에 해당할 것 같다.

결론적으로 두 tagging은 유사한 성능을 보이며, 성능 차이는 크게 없다. 이 pos-tagger들이 실제와 얼마나 똑같은지는 다른 문제이겠지만, 지금까지의 연구결과를 보았을 때 꽤 높은 수준의 성능을 보여주는 것 같다.